



## OPEN Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain

Agnieszka Pluta<sup>1,2,✉</sup>, Joanna Mazurek<sup>1</sup>, Jakub Wojciechowski<sup>2,3</sup>, Tomasz Wolak<sup>2</sup>, Wiktor Soral<sup>1</sup> & Michał Bilewicz<sup>1</sup>

The widespread ubiquity of hate speech affects people's attitudes and behavior. Exposure to hate speech can lead to prejudice, dehumanization, and lack of empathy towards members of outgroups. However, the impact of exposure to hate speech on empathy and propensity to attribute mental states to others has never been directly tested empirically. In this fMRI study, we examine the effects of exposure to hate speech on neural mechanisms of empathy towards ingroup (Poles) versus outgroup members (Arabs). Thirty healthy young adults were randomly assigned to 2 groups: hateful and neutral. During the fMRI study, they were initially exposed to hateful or neutral comments and subsequently to narratives depicting Poles and Arabs in pain. Using whole-brain and region of interest analysis, we showed that exposure to derogatory language about migrants attenuates the brain response to someone else's pain in the right temporal parietal junction (rTPJ), irrespective of group membership (Poles or Arabs). Given that rTPJ is associated with processes relevant to perspective-taking, its reduced activity might be related to a decreased propensity to take the psychological perspective of others. This finding suggests that hate speech affects human functioning beyond intergroup relations.

The spread of hate speech, a derogatory language based on racial, religious, or sexual prejudice, has become one of the key challenges for contemporary societies<sup>1–3</sup>. According to the epidemic model of hate speech<sup>4</sup>, people exposed to hate speech in their environment become prone to use derogatory language and engage in other forms of intergroup discrimination due to normative, emotional, and behavioural transitions caused by frequent exposure to such language.

As repeated exposure to violence reduces people's emotional responsiveness and decreases empathy when observing subsequent violent actions<sup>5–8</sup>, it is plausible that a similar process could be observed among people continuously exposed to verbal violence<sup>6,7</sup>.

Indeed, initial survey and experimental results<sup>9</sup> suggest that hate speech leads to desensitization, and this in turn affects intergroup relations. Conversely, recent findings suggest that empathy-inducing interventions can successfully resensitize individuals to hate speech and make them less supportive of the use of such statements<sup>10</sup>.

Empathy is a multifaceted construct referring to the human's capacity to share and understand others' feelings that are induced by the perception of another person's affective state e.g. pain<sup>11</sup>. Its elicitation is modulated by many interrelated factors<sup>12</sup>. For example, lack of empathy toward outgroup members has been observed in several empirical studies concerning intergroup empathy bias<sup>13,14</sup> and parochial empathy<sup>15,16</sup>. Although these phenomena seem to be fundamental aspects of intergroup relations, it is possible that specific social pressures could enhance them, leading to empathic failure<sup>17</sup>. Hate speech could be considered one of the factors enhancing empathic failure in intergroup relations.

Another stream of research on intergroup interaction has focused on the ability to attribute mental states—beliefs, intents, desires, emotions, and knowledge—to ourselves and others, referred to as theory of mind (ToM) or mentalizing<sup>18</sup>. Studies on dehumanized perception found that people watching members of groups that are considered lacking both warmth and competence react with lower mentalization<sup>19–21</sup>. This suggests that a specific

<sup>1</sup>Faculty of Psychology, University of Warsaw, Stawki 5/7 Street, 00-183 Warszawa, Poland. <sup>2</sup>Bioimaging Research Center, World Hearing Center of Institute of Physiology and Pathology of Hearing, Warszawa, Poland. <sup>3</sup>Laboratory of Emotions Neurobiology, Nencki Institute of Experimental Biology PAS, Warsaw, Poland. ✉email: apluta@psych.uw.edu.pl

form of hate speech that dehumanizes certain groups could particularly deteriorate the ability to understand mental states of others.

Lastly, the overexposure to aggression (in e.g. media) and suffering of others may desensitize people to violence as well as to pain of others, leading to compassion collapse—the condition which is marked by psychic fatigue, reduced empathy and feeling detached with others<sup>8,22–24</sup>. Applying this model to hate speech, one could also expect that exposure to such forms of verbal violence could deteriorate people's ability to empathize and/or mentalize resulting in a more general empathic numbing—a state of depletion in which compassion cannot be generated towards other people.

These three perspectives offer different predictions about the potential effects of exposure to hate speech on the ability to share and understand the mental states of others. According to the empathic failure model, hate speech should enhance existing biases in empathy by increasing empathy toward ingroup members and reducing outgroup-directed empathy. According to the dehumanized perception model, hate speech should reduce the mentalization propensity toward dehumanized outgroups, but should not affect empathy towards ingroup members. According to the empathic numbing hypothesis, exposure to hate speech should limit the ability to empathize (or/and mentalize) with other people, regardless of their group membership.

The behavioral and self-report character of previous studies did not allow the identification of the specific neurocognitive mechanism that would link empathy and/or mentalizing abilities to the use of hate speech. In this study, we seek to answer this question using a social neuroscience approach and functional magnetic resonance imaging (fMRI) techniques.

Although both empathy and ToM are multidimensional and often intertwined abilities that are jointly required in many social situations, they are rooted in separable brain networks<sup>25</sup>. ToM involves an extended neural network (collectively referred to as the ToM network) comprising the temporoparietal junction (TPJ), precuneus (PC), posterior superior temporal sulcus (pSTS), medial prefrontal cortex (MPFC), and temporal poles<sup>26–28</sup>. Among them, the TPJ (especially in the right hemisphere) is considered to play a critical role in processing information about someone else's psychological perspective. Studies that have examined the neural basis of empathy have identified regions, such as the anterior cingulate (ACC) cortex<sup>29</sup> and anterior insula (AI)<sup>30</sup> (collectively referred to as the pain matrix) that may be essential for processing information about vicarious experiences of other people. An accumulating body of evidence shows that the ACC and the AI respond to both first- and third-person experiences of pain. Therefore, this response pattern might indicate “state-matching” between the observer and target<sup>31</sup>.

Importantly, a number of studies have shown that people process information about the affective and cognitive mental states of members of distinct ethnic or cultural groups differently<sup>32</sup>. For example, it has been demonstrated that people show diminished brain activity of the ACC in response to physical pain experienced by others of different ethnicities<sup>33</sup>. In contrast, the study of Marthur and colleagues<sup>34</sup> demonstrated that while confronted with the suffering of the victims of hurricane Katrina, African-Americans and Caucasian-Americans differently activated the mentalizing network (specifically MPFC) but not the pain matrix. The brain response varied depending on the victim's ethnicity.

These neuropsychological differences in reactions to ingroup and outgroup members' emotional or physical suffering may be further amplified in contexts dominated by hate speech. Surprisingly, no studies we know have directly examined the impact of exposure to hateful statements (hate speech) on the neurocognitive processes.

In the present study, we aimed to disentangle three potential consequences of exposure to hate speech (e.g. the empathic failure, the dehumanized perception, the empathic numbing) by examining neurocognitive mechanisms of empathy and ToM. To this end, we conducted a study of Polish individuals and chose Arabs as the outgroup because in recent times, this group has been subject to political attacks in the Polish media<sup>35</sup>. While in the magnetic resonance 3 T scanner, individuals were presented with long lists of negative online comments about the outgroup (hateful condition) or not (neutral condition), interspersed with painful and non-painful stories featuring protagonists of Polish (ingroup) and Arab (outgroup) ethnicities.

Therefore, we aimed to test the following hypotheses:

H1. According to the empathic failure model, exposure to hate speech leads to decreased brain activity within the AI and ACC in response to outgroup members' pain or suffering, and increased activity within these areas in response to ingroup members' pain or suffering.

H2. According to the dehumanized perceptions model, exposure to hate speech leads to attenuation of the ToM network in response to outgroup (but not ingroup) members' pain or suffering.

H3. According to the empathic numbing hypothesis, exposure to hate speech leads to decreased activity in the pain matrix and/or ToM network in response to others' pain or suffering, regardless of their group membership.

## Method and materials

**Participants.** Thirty right-handed Polish participants ( $M_{age} = 23.7$ ,  $SD = 4.28$ , females = 18;  $M_{education\ years} = 14.9$ ,  $SD = 2.10$ ) with no reported history of neurological or psychiatric diagnoses, having no contraindications to taking part in magnetic resonance imaging examinations, volunteered to participate in the study from April 2019 to August 2019. Twenty seven participants had declared their ethnic identity as only Polish, 2 described their ethnicity as Polish-American, and 1 as Polish-Japanese. All subjects were born and raised in Poland. They had normal or corrected-to-normal vision, provided written informed consent, and were compensated with a small gift. The experimental procedure was approved by the Ethical Committee of the Faculty of Psychology, University of Warsaw, and was carried out in accordance with the Declaration of Helsinki.

**Procedure.** Participants were randomly assigned to two equinumerous groups ( $n = 15$  each), each with equal gender proportion (9 women and 6 men), and exposed either to negative comments about the Arabs (hateful) or

about current social issues (neutral). The study had a mixed 2 (Type of Comments: hateful vs. neutral)  $\times$  2 (Type of Stories: pain vs. no pain)  $\times$  2 (Protagonist's Ethnicity: Polish vs. Arab) design. The Type of Comments was a between-subject factor, while the other two (Stories and Protagonist's Ethnicity) were within-subject factors. Sensitivity analysis was conducted with G\*Power to find a minimum detectable effect (MDE) for a 2 (between)  $\times$  4 (within) interaction (with power = 0.80 and  $\alpha$  = 0.05). These yielded a MDE value equivalent to Cohen's  $f$  = 0.22 or  $\eta^2$  = 0.045 (see the power analysis protocol uploaded on the OSF repository: <https://osf.io/t8fra>).

**Study design.** The whole study consisted of two parts: a self-report measure and fMRI examination (described in the sections below). During the fMRI part, we used an experimental manipulation in the form of internet-like comments (hate speech or neutral), followed by the Narrative-based Pain Empathy Task (described in the sections below). To maintain the impact of the comments during the cognitive task, we used short, consecutive manipulations followed by the task runs repeated four times. The study design is presented in Fig. 1.

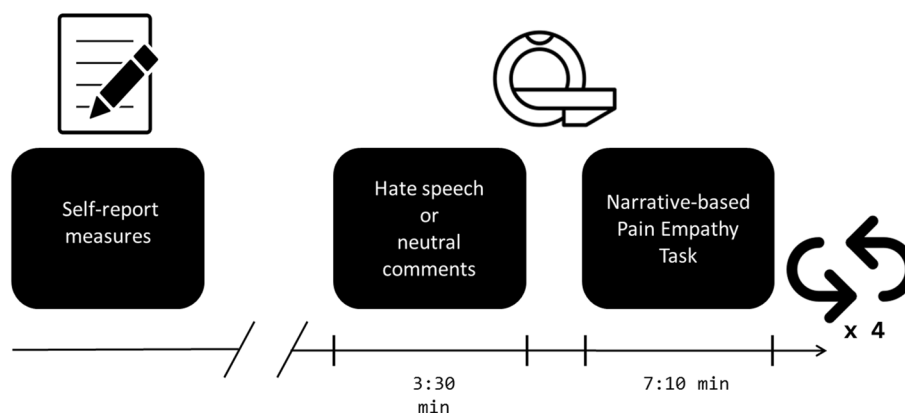
**Self-report measures.** Two to seven days before fMRI testing, participants completed a questionnaire (administered through Qualtrics Survey Software; Qualtrics, Provo, UT) including demographic variables and two measures of political ideology (right-wing authoritarianism and social dominance orientation) that are known correlates of hate speech attitudes<sup>36</sup>. Right-wing authoritarianism was measured with 6 items (such as "Obedience and respect for authority are the most important values children should learn") of a shortened version of Funke's (2005) the three-dimensional Right-Wing Authoritarianism (RWA) scale ( $\omega_t$  = 0.86,  $M$  = 3.42,  $SD$  = 1.25). The Polish adaptation of the Social Dominance Orientation (SDO) scale<sup>37</sup>, shortened to eight items (such as "Superior groups should dominate inferior groups"), was used to assess participants' preference for social hierarchy over egalitarianism ( $\omega_t$  = 0.92,  $M$  = 2.86  $SD$  = 1.18). The responses to RWA and SDO were coded on a 7-point scales from 1 ("Definitely disagree") to 7 ("Definitely agree").

**Exposure to hateful or neutral comments.** In the first part of the procedure, participants in the scanner were exposed to comments (hateful or neutral). The comments were selected based on previous studies on exposure to online hate speech<sup>9</sup> and were presented in a form imitating real-life internet forums. The comments' selection procedure was described in detail in the work by Bilewicz and colleagues<sup>38</sup>.

In the hateful comments condition, the comments included derogatory statements directed against Arab and Muslim immigrants and refugees (e.g., "It is high time. This wall should have been put up a long time ago. We need to defend Europe against the hordes"). In the neutral comments condition, the comments focused on current economic and legal issues, as well as international and domestic politics (e.g., "Is it just me who gets pissed off by this neologism: prosumer? Who came up with something so dumb, and why is everyone using it?"). In both conditions, the affective tone of the comments varied from mildly negative to offensive, but the target was different. Note that the label "neutral" means that the comments were not directed against the Arab outgroup (nor any minority group), and not that they were emotionally neutral. The sample comments page can be seen in Supplementary Information Fig. S1. The full list of comments is available on the Open Science Framework (Supplemental Materials A <https://osf.io/nw8xq>).

Each block in this part consisted of 5 pages of comments (on average three per page) displayed in random order, each for 30 s. A fixation cross (+) was displayed for 20 s in between the comments (C) in the following order: CC + CC + C+. Participants were instructed to read the comments that appeared on the screen and to try to remember them as well as possible. Overall each commentary run lasted 3:30 min. Because we had only 30 unique comments divided into 10 pages, each comment was displayed twice (before I and III as well as II and IV runs of the Narrative-based Pain Empathy Task).

**Narrative-based Pain Empathy Task.** After each run of comments, subjects performed a Narrative-based Pain Empathy Task. There were 64 narrative stories used in the study—32 pain stories (describing painful physical sensations e.g. protagonist cut his finger with a sharp knife) and 32 no painful versions of each story. The



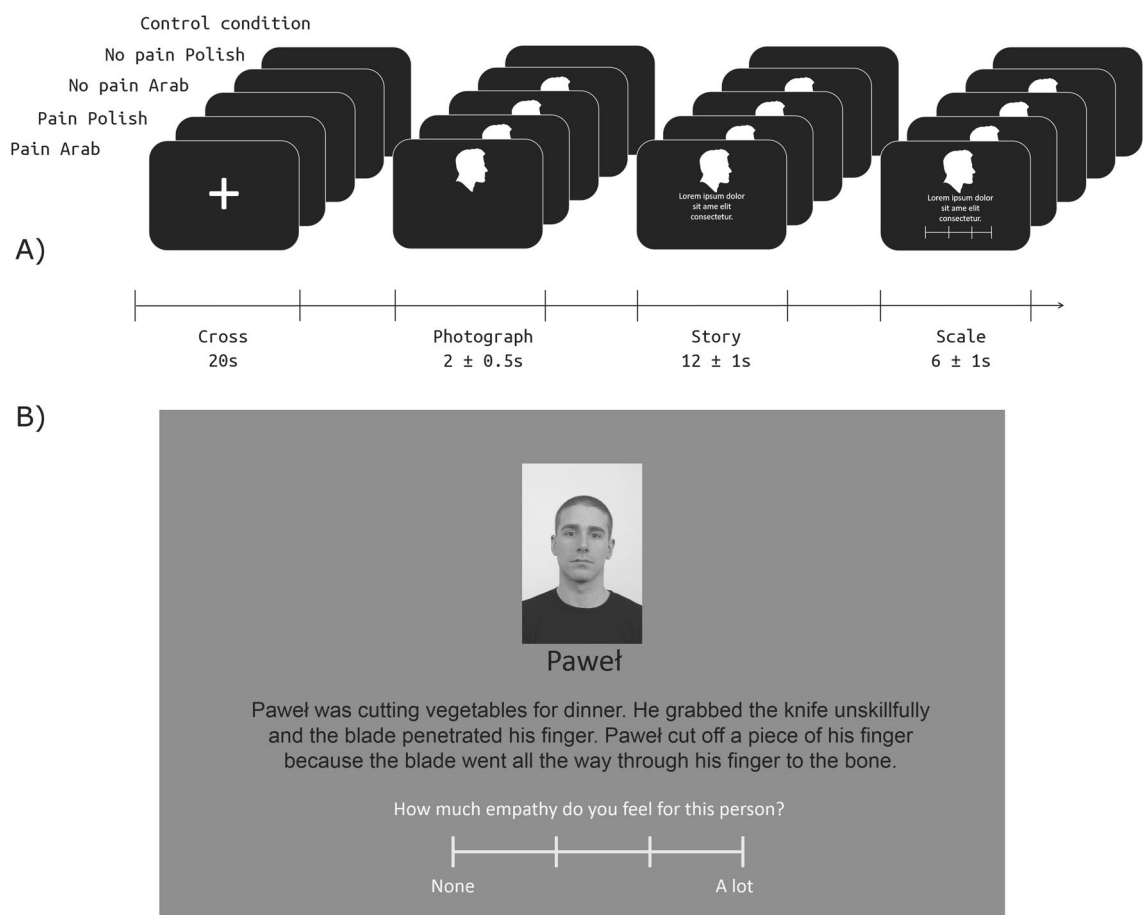
**Figure 1.** The design of the study.

stories were based on Bruneau and colleagues' experiment (2012) and translated into Polish. The original version and the back-translated versions were compared by a native speaker of Polish, considering the comparability of language and similarity of interpretability. Each of the stories was repeated twice, once with a Polish name and the corresponding picture (above the text) of a protagonist, and the other with an Arab name and the picture of the protagonist<sup>39</sup>. The full list of stories is available on the OSF (Supplemental Materials B <https://osf.io/fcnuy/>).

The pictures used in the experimental task were faces from the Radboud Faces Database<sup>40</sup>. The database contains 18 Moroccan adult male faces, which were chosen to represent Arab protagonists, and 20 Caucasian adult male faces, which were chosen to represent Polish protagonists. A sample picture can be seen in Supplementary Information Fig. S2.

Participants were instructed to attentively read the stories accompanied by the names and faces (Photograph). Each story was displayed for 20 s on average (Story). Before each story, a fixation cross was presented for 20 s (Cross). Then, a face picture with the protagonist's name underneath was displayed for  $2 (\pm 0.5 \text{ s jitter})$  seconds on its own, then the story appeared underneath for an additional 12 s ( $\pm 1 \text{ s jitter}$ ). At the end, a response prompt with a scale appeared underneath the story for 6 s ( $\pm 1 \text{ s jitter}$ ) (Scale). A schematic display of fMRI design with example items of stimuli and a sample page that participants saw in the scanner can be seen in Fig. 2.

After each stimulus, there was an interstimulus interval (ISI) on average for 2 s ( $\pm 0.5 \text{ s jitter}$ ). The prompt asked participants to indicate their empathy towards the protagonist on the following scale: 1 (*none*) to 4 (*a lot*). Participants responded using an MR-safe button box. Average responses and reaction times (RTs) for each condition were computed for each individual. Each run consisted of 16 stories (4 per condition) and 4 fixation crosses, displayed in a pseudorandom order so that no two stimuli of the same type (pain Polish, pain Arab, no pain Polish, no pain Arab, fixation cross) were shown in a row. For each run, a sequence was chosen randomly, and for each stimulus type, the stimuli were assigned randomly without repetition. The whole Narrative-based Pain Empathy Task consisted of 4 runs each lasting 7:10 min. Therefore, the whole fMRI experiment lasted 42:40 min. Stimulus presentation was controlled through a script written in Presentation 21.0 software (Neurobehavioral Systems Inc.) and displayed on MR-compatible full HD display screen (NordicNeuroLab Inc.) via a mirror mounted on the head coil.



**Figure 2.** (A) A schematic display of the design of the Narrative-based Pain Empathy Task (N-PET) in fMRI. Each experiment condition had the same block structure. In the control condition, the fixation cross was displayed. (B) A sample stimulus displayed in the N-PET at the final phase (Scale), when participants were instructed to indicate their empathy towards the protagonist.

At the end of the study, the participants were fully debriefed regarding the nature of the study and compensated for their participation with small gifts. Finally, the participants were asked open-ended questions about the Hateful or Neutral Comments to make sure that they had been paying attention to them. Based on these answers, we decided that all participants paid attention to the stimuli and could be included in the further analysis.

**MRI data acquisition.** The study took place in the Bioimaging Research Center, Institute of Physiology and Pathology of Hearing, Kajetany, Poland. MRI scanning was performed on a 3 T Siemens Prisma MRI scanner equipped with a 20-channel phased-array RF head coil. Functional data for all tasks were acquired using a Multi-band (Simultaneous Multi-Slice) echo-planar-imaging (EPI) sequence (TR = 1500 ms, TE = 27 ms, flip angle = 90°, FOV = 192 × 192 mm, 94 × 94 mm image matrix, 48 transversal slices of 2.4 mm slice thickness, voxel size of 2.0 × 2.0 × 2.4 mm, Slice Accel. Factor = 2, In-Plane Accel. Factor = 2, IPAT = 4, TA = 7:10 min).

Structural images were collected with a T1-weighted 3D MP-Rage sequence (TR = 2300 ms, TE = 2.26 ms, TI = 900 ms, 8° flip angle, FOV = 208 × 230 mm, image matrix 232 × 256 mm, voxel size of 0.90 × 0.90 × 0.90 mm, 208 slices of 0.90 mm slice thickness, TA = 4:53 min).

**Image processing.** Preprocessing and first-level analysis of the fMRI data was carried out with the use of SPM12 (SPM; WellcomeTrust Centre for Neuroimaging, London, UK). The functional data were slice-time corrected, motion corrected, realigned to the first image from each run and co-registered to the individual structural images. High-resolution structural images were segmented and normalized to the common MNI space with resampling to 1 mm isometric voxels. The obtained transformation parameters were applied to the functional volumes with resampling to 2 mm isometric voxels. Spatial smoothing with a Gaussian kernel of full-width half-maximum (FWHM) of 6 mm was performed on the normalized functional images. Additional high-pass filters of 256 s were used for the functional data.

For the first-level statistical analysis, a general linear model was created including a regressor for each condition (pain Polish, pain Arab, no pain Polish, no pain Arab, fixation cross). This resulted in 5 regressors in total. Additionally, six movement parameters (3 translations, 3 rotations) were included as nuisance regressors. The conditions were modeled using a standard boxcar model. For the second-level analysis, we used the Multivariate and Repeated Measures (MRM) toolbox<sup>41</sup>.

To test for a simple effect of pain, at the second level, a two-sample *t*-test was performed using the (pain Arab + pain Polish) > (no pain Arab + no pain Polish) contrast for all participants (*N* = 30). For interaction analysis, images of beta estimates of the experimental conditions from each participant were entered into a three-way repeated measures ANOVA with one between-subject factor: Type of Comments (neutral vs. hateful) and two within-subject factors: Protagonist's Ethnicity (Polish vs. Arabs) and Story Type (pain vs. no pain). Finally, for follow-up post-hoc analysis, voxels from each cluster that showed significant effect/interaction were averaged and then tested with appropriate post-hoc contrasts.

All reported contrasts were thresholded voxel-wise at *p* < 0.001 and family-wise error (FWE) corrected at the cluster level  $p_{\text{FWE}} < 0.05$ . Bspmview toolbox, based on the Anatomy Toolbox, was used for automated anatomical labeling of the results (<https://github.com/spunt/bspmview>). For visualization of results in surface-based form, BrainNet Viewer was used<sup>42</sup>.

Subsequently, we conducted Regions of Interest (ROI) analysis based on regions postulated to be involved in empathy, mentalizing, and moral sensitivity in relation to outgroup bias<sup>32</sup> (AI, ACC, TPJ and mPFC). Neurosynth (Neurosynth.org) with the term “empathy” was queried in order to identify the peak coordinates of the preselected ROIs. This resulted in images from 187 PubMed publications within the regions comprising both the classical “pain matrix” and “ToM matrix” as well brain regions beyond these networks. Finally we created 8 mm spheres centered around the peak coordinates of the left AI [−34, 22, 0], right AI [38, 22, −8], left dorsal ACC (dACC) [−6, 20, 44], right dACC [10, 24, 30], left TPJ [−48, −62, 20], and right TPJ [50, −50, 18], mPFC [2, 24, 56]. ROIs are presented in Fig. 3.

The averaged beta estimates for each condition of each task from the ROIs were extracted using in-house Matlab scripts.

First, we conducted paired *t*-tests analysis for pain and no pain conditions for the right and left AI and dACC for each group separately to ensure that painful stories elicited greater BOLD response in the key regions of empathy for physical pain irrespectively of experimental manipulation (neutral vs. hateful).

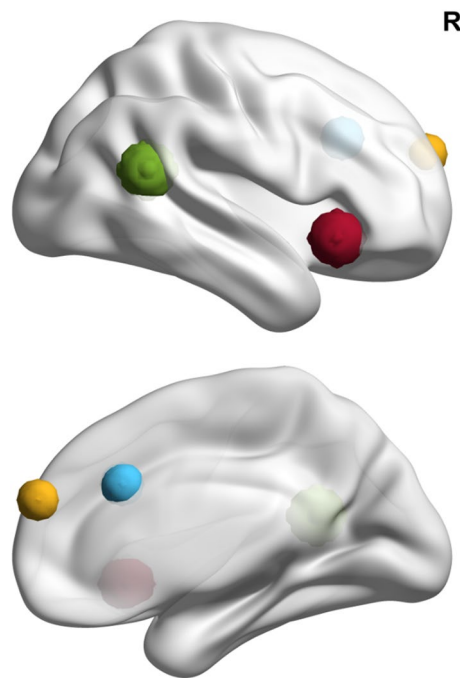
Finally, a series of ANOVAs was conducted on the following ROIs (selected a priori, see Methods): left and right TPJ, mPFC, left and right insula, and left and right dACC. Each ANOVA followed the same full factorial mixed design, with Type of Comments (neutral vs. hateful, between-subject), the Story Type (no pain vs. pain, within-subject), and Protagonist's Ethnicity (Arab vs. Polish).

The analyses were carried out using R software(<http://www.rstudio.com/>).

## Results

**Self-report results obtained during fMRI study.** Ratings of empathy towards protagonists in no pain (vs. pain) were low for all participants (the average of median ratings was *M* = 1 and the standard deviation was *SD* = 0). Conversely, average empathy ratings toward protagonists in pain (vs. no pain) were significantly larger than the midpoint of the scale which was 2.5 (*M* = 3.39, *SD* = 0.69), *t*(29) = 7.04, *p* < 0.001, *d* = 1.29. Therefore, to avoid issues due to the lack of homogeneity of variances, further analyses in this section were conducted only on ratings toward protagonists in pain. Median empathy ratings for each participant were transformed using the aligned rank method<sup>43</sup>, and then analyzed with factorial mixed-design ANOVA, with the Type of Comments (neutral vs. hateful) as a between-subject factor, and the Protagonist's Ethnicity (Arab vs. Polish) as a within-subject factor. Participants exposed to hateful comments (*M* = 3.23, *SD* = 0.75) empathized to a slightly





**Figure 3.** Regions of Interest selected based on Neurosynth survey (only ROIs in the right hemisphere are visualized, ROIs in the left hemisphere were roughly symmetrical, see “fMRI analysis” section for details). AI—red, TPJ—green, dACC—blue, mPFC—yellow.

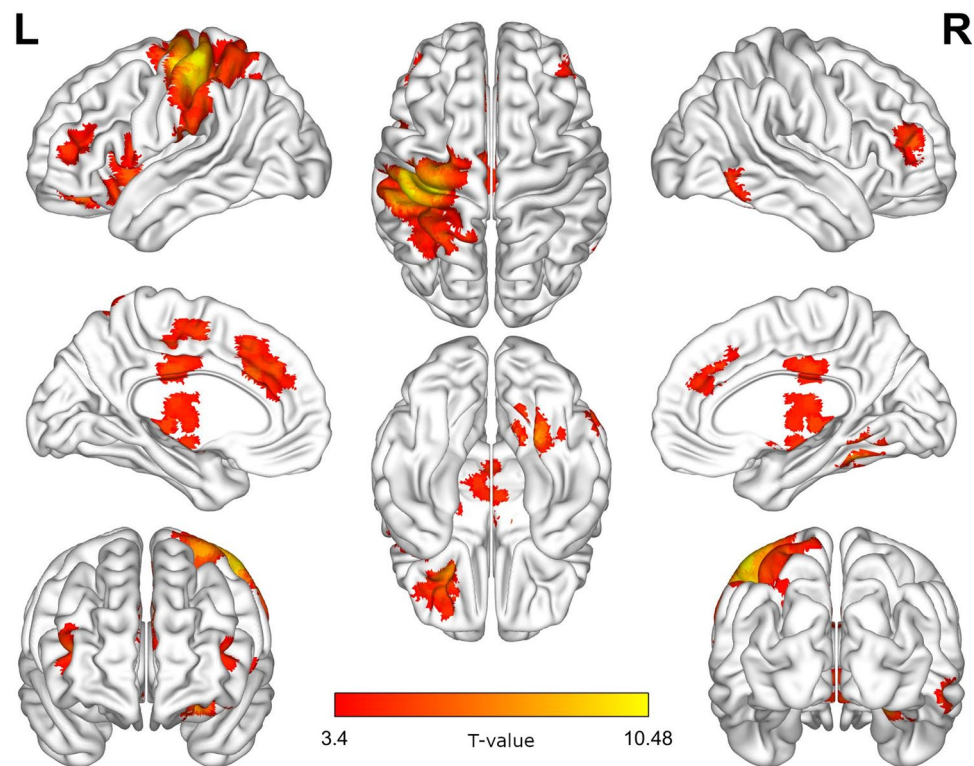
lesser extent than participants exposed to neutral comments ( $M = 3.55$ ,  $SD = 0.63$ ). However, the difference was not statistically significant,  $F(1, 28) = 1.62$ ,  $p = 0.213$ ,  $\eta_p^2 = 0.055$ . Participants were also more empathic to Polish ( $M = 3.42$ ,  $SD = 0.68$ ) than to Arab protagonists ( $M = 3.37$ ,  $SD = 0.74$ ), but also in this case the difference was not statistically significant,  $F(1, 28) = 0.95$ ,  $p = 0.337$ ,  $\eta_p^2 = 0.033$ . The interaction of the Type of Comments and the Protagonist's Ethnicity was not significant,  $F(1, 28) = 0.62$ ,  $p = 0.438$ ,  $\eta_p^2 = 0.022$ . In a similar vein, analyses of log-transformed RTs in mixed design ANOVA (Type of Comments  $\times$  Protagonist's Ethnicity) revealed no significant effect or interactions. Specifically, the main effect of Type of Comments on RT was not significant,  $F(1, 28) = 2.92$ ,  $p = 0.099$ ,  $\eta_p^2 = 0.090$ . The main effect of Protagonist's Ethnicity was not significant,  $F(1, 28) = 1.07$ ,  $p = 0.310$ ,  $\eta_p^2 = 0.040$ . The Type of Comments  $\times$  Protagonist's Ethnicity interaction was not significant,  $F(1, 28) = 0.002$ ,  $p = 0.963$ ,  $\eta_p^2 < 0.001$ .

**Functional MRI results.** *Whole-brain analysis.* The analysis revealed the main effect of Story Type (no pain vs. pain, within-subject). Stories with Pain elicited higher blood oxygen level dependent (BOLD) signal involved in processing information about pain experienced by another person (including the affective components): AI, ACC, pre- and postcentral gyrus, thalamus, caudate, supplementary motor cortex, orbitofrontal cortex, frontal pole bilaterally, cerebellar vermis (4/5). All regions visualized in Fig. 4 are displayed in Table 1.

In addition, there was also a significant Type of Comments  $\times$  Story interaction in the right TPJ, which is a hub of mentalizing networks (Fig. 5). Interestingly, post-hoc analysis for this region showed that exposure to hateful comments attenuated its activity.

*Region of interest analysis.* Paired t-test analysis yielded that in both groups (hate, neutral) stories with pain were associated with greater BOLD response than stories without pain in the left dACC (hate group:  $M_{\text{pain}} = 0.51$ ,  $SE_{\text{pain}} = 0.27$  vs.  $M_{\text{nopain}} = -0.04$ ,  $SE_{\text{nopain}} = 0.23$ ,  $t(14) = 3.66$ ,  $p < 0.05$ ,  $d = 0.94$ ; neutral group:  $M_{\text{pain}} = 0.73$ ,  $SE_{\text{pain}} = 0.28$  vs.  $M_{\text{nopain}} = 0.48$ ,  $SE_{\text{nopain}} = 0.22$ ,  $t(14) = 3.28$ ,  $p < 0.05$ ,  $d = 0.85$ ). In addition, participants exposed to hateful messages reacted with increased BOLD response in right AI ( $M_{\text{pain}} = 1.01$ ,  $SE_{\text{pain}} = 0.38$  vs.  $M_{\text{nopain}} = 0.59$ ,  $SE_{\text{nopain}} = 0.36$ ,  $t(14) = 2.72$ ,  $p < 0.05$ ,  $d = 0.7$ ) and participants exposed to neutral comments reacted with higher BOLD response in left AI ( $M_{\text{pain}} = 0.34$ ,  $SE_{\text{pain}} = 0.18$  vs.  $M_{\text{nopain}} = 0.16$ ,  $SE_{\text{nopain}} = 0.16$ ,  $t(14) = 3.18$ ,  $p < 0.05$ ,  $d = 0.67$ ) when exposed to pain experienced by the protagonist as compared to no pain stories. This suggests that in both groups the presentation of painful stories resulted in higher BOLD activation in key regions implicated in empathy for physical pain.

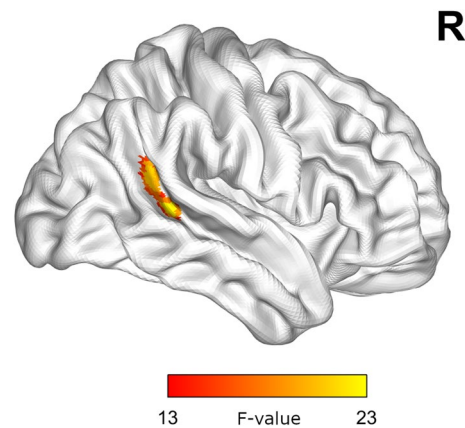
The analyses of ANOVAs conducted on 7 ROIs revealed no significant effects of Type of Comments or Protagonist's Ethnicity. However, analyses revealed significant main effects of Story. Stories with pain were associated with greater BOLD response than stories without pain in the left insula ( $M = 0.18$ ,  $SE = 0.13$  vs.  $M = 0.05$ ,  $SE = 0.14$ ), the right insula ( $M = 0.94$ ,  $SE = 0.23$  vs.  $M = 0.70$ ,  $SE = 0.21$ ), the left dACC ( $M = 0.62$ ,  $SE = 0.18$  vs.  $M = 0.22$ ,  $SE = 0.16$ ), and the right dACC ( $M = 0.33$ ,  $SE = 0.15$  vs.  $M = 0.16$ ,  $SE = 0.15$ ). Conversely, stories with pain were associated with decreased BOLD response in comparison to stories without pain in the left TPJ ( $M = -0.10$ ,  $SE = 0.27$  vs.  $M = 0.31$ ,



**Figure 4.** Brain activations (T-map) showing a significant main effect of pain by contrasting conditions: (pain Polish + pain Arab) > (no pain Polish + no pain Arab). Thresholded for voxel-level at  $p < 0.001$  and FWE-corrected  $p < 0.05$  for cluster size.

Location	Cluster	T-value	Coordinate (MNI)		
			X (mm)	Y (mm)	Z (mm)
Postcentral gyrus L	4901	10.4812	−46	−28	62
Precentral gyrus L	4901	7.6525	−30	−12	68
Superior parietal gyrus L	4901	5.903	−30	−50	64
Lobule IV, V of cerebellar hemisphere R	1242	9.4459	18	−48	−22
Lobule IV, V of vermis	1242	5.5222	4	−62	−10
Crus II of cerebellar hemisphere R	1242	4.5072	54	−48	−40
Caudate nucleus R	872	8.1566	12	14	6
Ventral posterior nucleus of thalamus L	872	6.2541	−12	−20	8
Posterior orbital gyrus L	759	7.4245	−26	28	−14
Insula L	759	5.88	−40	10	−4
Middle frontal gyrus L	759	4.6561	−24	52	−12
Anterior cingulate cortex, subgenual L	521	6.1257	−8	24	30
Anterior cingulate cortex, pregenual R	521	4.6454	4	40	24
Middle frontal gyrus R	220	5.7797	40	44	14
Inferior temporal gyrus R	166	5.5933	60	−64	−10
Middle cingulate L	266	5.5132	−2	−22	32
Supplementary motor area L	266	5.107	−6	−6	54
Middle frontal gyrus L	205	5.1466	−36	36	18
Crus I of cerebellar hemisphere L	189	4.8852	−34	−70	−28

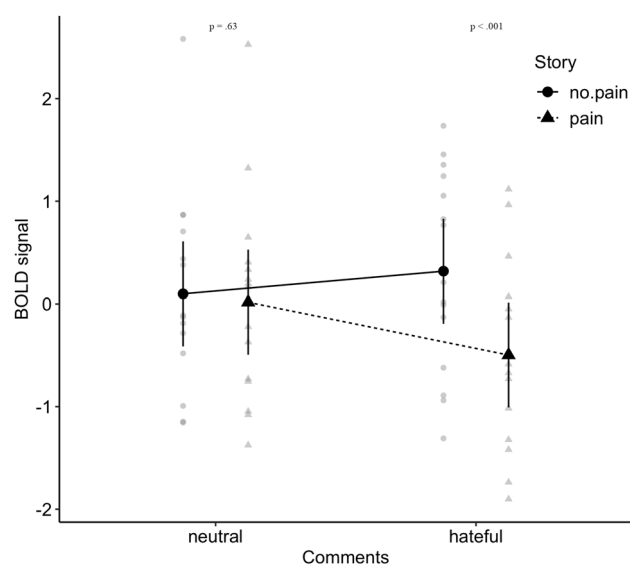
**Table 1.** Regions showing a significant main effect of pain.



**Figure 5.** Brain activations (F-map) showing a significant interaction of Type of Comments  $\times$  Story effect. Thresholded for voxel-level at  $p < 0.001$  and FWE-corrected  $p < 0.05$  for cluster size.

Effect	Left TPJ		Right TPJ		Left insula		Right insula		Left dACC		Right dACC	
	F	$\eta_p^2$	F	$\eta_p^2$	F	$\eta_p^2$	F	$\eta_p^2$	F	$\eta_p^2$	F	$\eta_p^2$
C	0.92	0.028	0.20	0.006	0.11	0.003	0.00	<0.001	1.30	0.039	0.83	0.025
E	0.18	<0.001	1.08	0.001	0.00	<0.001	0.05	<0.001	0.79	0.001	0.07	<0.001
C $\times$ E	0.03	<0.001	0.46	<0.001	0.44	0.001	0.87	<0.001	1.40	0.002	0.46	<0.001
S	7.36*	0.018	14.52***	0.050	6.24*	0.008	6.11*	0.010	22.49***	0.045	5.95*	0.010
C $\times$ S	0.45	0.001	9.75**	0.034	3.29	0.004	3.48	0.006	3.32	0.007	0.10	<0.001
E $\times$ S	0.03	<0.001	0.09	<0.001	3.41	0.007	3.97	0.002	1.62	0.001	3.33	0.003
C $\times$ E $\times$ S	1.52	0.002	1.65	0.003	0.02	<0.001	0.00	<0.001	0.83	<0.001	0.65	<0.001

**Table 2.** Results of mixed design analyses of variance for selected regions of interest. N = 30. Degrees of freedom are 1 and 28 for all effects. C = Comments (neutral vs. hateful), E = Ethnicity (Arab vs. Polish), S = Story (no pain vs. pain). \* $p < 0.05$  \*\* $p < 0.01$  \*\*\* $p < 0.001$ .



**Figure 6.** Type of Comments (Neutral vs. Hateful)  $\times$  Story (No Pain vs. Pain) Interaction Observed in the Right TPJ. Black points indicate estimated means. Gray points indicate individual observations. Vertical lines indicate 95% confidence intervals.



$SE=0.28$ ) and right TPJ ( $M=-0.24$ ,  $SE=0.18$  vs.  $M=0.21$ ,  $SE=0.18$ ). However, the effect of the story on the right TPJ was qualified by the significant Type of Comments  $\times$  Story interaction (see Table 2). As shown in Fig. 6, the difference in signal in the right TPJ between pain and no pain stories observed was significant among participants exposed to hateful comments,  $t(28)=-4.90$ ,  $p<0.001$ ,  $d=-0.81$ , but not significant among participants exposed to neutral comments,  $t(28)=-0.49$ ,  $p=0.63$ ,  $d=-0.08$ . Thus, participants exposed to hateful messages reacted with reduced right TPJ activity when exposed to pain experienced by the protagonist as compared to no pain stories. This suggests that among participants exposed to hateful (but not neutral comments), painful stories resulted in lower activation of the right TPJ, irrespective of the protagonist's ethnicity.

None of the remaining two-way or three-way interactions were significant. For mPFC none of the effects were statistically significant. In all, ROI analysis (on regions identified based on the query in Neurosynth) was consistent with the whole brain analysis.

## Discussion

This experimental study tested the effects of prior hate speech exposure on neurocognitive mechanisms of empathic reaction and/or propensity to attribute mental states to others when contemplating another's pain in an intergroup setting. Building our predictions on previous research on intergroup bias in empathy, as well as desensitization to verbal aggression, in this study, we investigated 3 possible theoretical models: the empathic failure model, the dehumanized perception model, and the empathic numbing model.

The results demonstrate that exposure to hate speech, even for about a quarter of an hour, affects brain response by reducing BOLD response in the region involved in mentalizing (rTPJ) while faced with the pain of others, regardless of the group membership of the protagonist (Polish or Arab).

Firstly, whole-brain analysis revealed that the narrative scenarios used in the fMRI experiment activate brain regions typically considered as the core empathy network: AI, ACC, SMA, pre- and postcentral gyrus, parietal lobe, thalamus, frontal pole (specifically the inferior frontal gyrus)<sup>44</sup>. The ROI analysis was in line with the whole-brain results and showed increased activity within the insula and dACC bilaterally in response to painful stimuli. These results confirm that stimuli depicting another person's pain lead to the recruitment of the pain matrix regions as planned.

Secondly, neither fMRI nor self-report measures collected during the fMRI session were indicative of in-group biases. On the one hand, the lack of intergroup bias in empathy ratings may be related to the fact that self-report assessment during the fMRI study was done on a 4-point scale, which may not have been sensitive enough to capture possible subtle ingroup bias. These results are also consistent with the study of Bruneau and colleagues<sup>45</sup>, which found that being witness to physical pain experienced by distant outgroups evokes the same level of compassion, as well as similar neural activity at the whole-brain level within the pain matrix, as observing the pain of members of one's own group. Given that Poles have not been involved in ongoing conflicts with Arabs, do not have a history of conflict with them, and that the two groups come from very different geographic, linguistic, and political backgrounds, Arabs conform to the definition of a distant outgroup for Poles.

This result might be also explained by the participants' sociodemographic characteristics (young well-educated adults who have had an average of 14.9 years of schooling) and ideological orientation. Given that a preference for hierarchy has been shown to be associated with reduced empathy towards outgroup persons<sup>46</sup>, participants' moderate RWA and SDO scores could possibly explain the absence of parochial empathy (i.e., the tendency to show greater empathic concern for members of one's ingroup) in the studied group.

Finally, both whole-brain and independent ROI analysis revealed that exposure to derogatory language about migrants and minority groups suppresses the brain response of the rTPJ. This effect occurred regardless of the group membership of the protagonist (Polish or Arab). A sizable number of studies have shown the rTPJ to be robustly activated when people attribute mental states to others (intentions/beliefs), in order to explain or predict or make moral judgements about their actions<sup>47,48</sup>. Interestingly, it has also been demonstrated that disrupting one's rTPJ using TMS impairs the ability to consider one's beliefs and intentions when making moral judgments<sup>49</sup>. Therefore, reduction of rTPJ activity after exposure to hate speech might reduce the propensity to project ourselves into the reality of others in order to share/understand their psychological state. Given that the ability to adopt others' perspectives is an essential component of compassionate responses leading to altruistic behaviors<sup>50</sup>, the observed effect might have far-reaching consequences by hampering our responses towards needy people.

This result supports the psychological numbing hypothesis: repeated exposure to negatively charged messages with a high emotional intensity leads to psychological desensitization. Although this effect has not been demonstrated in relation to hate speech, it has been confirmed for other types of highly emotional stimuli, such as violent media stimuli and violent video-games. In particular, it has been shown that repeated exposure to media violence decreases negative affect<sup>51</sup>, playing violent games reduces physiological arousal (heart rate and galvanic skin response) in response to real-life violence<sup>52</sup>.

Although prosocial behavior was not measured in this study, given that previous research has shown that helping involves perspective-taking and is associated with increased activation of the mentalizing network (including the TPJ)<sup>53</sup>, it is plausible that reduced TPJ activation results in disengagement from inferring the cognitive and affective states of others and thus diminishes the propensity to help those in need. It has been demonstrated using an unobtrusive measure of empathy and mentalizing (i.e., neural processes involved in those phenomena), rather than self-report measures—the latter are subject to social desirability pressures, highly common in research on such sensitive issues as intergroup empathy. Further research is needed to test the behavioral consequences of this process.

**Limitations and future directions.** Numerous important theoretical questions remain for future research. Further research should resolve whether the observed effect of reduced activity of areas involved in

mentalization after exposure to hate speech is long-lasting. For instance, a longitudinal study has found that, despite numerous studies showing a decrease in empathy directly after playing video games, such effects do not persist over time<sup>54</sup>. If a similar relation was valid for negative media content and the decrease in brain activity within the mentalizing hub were transient, the consequences would not be as severe as in a scenario where the effects of such manipulation persist over time. Additionally, it would be interesting to ascertain the effect of political attitudes on patterns of neural activity after exposure to hate speech. As a next step, it would be worthwhile to examine whether, in fact, reduced mentalization decreases prosocial behavior when both in-group and out-group members experience hardship.

There are several limitations to this study. One of the limitations of our study is the sample size. Although it exceeds the median size for experimental fMRI studies in recent years<sup>55</sup>, it is still moderate in light of the between-subject study design used.

Moreover, the sample was not diverse in terms of political attitudes and education. To increase the generalizability of the results and mechanisms investigated, a more diverse sample should be examined. We also cannot rule out that, to some extent, the lack of ingroup bias may be related to low sensitivity in the experimental procedures. Due to time constraints, we chose to use only narratives describing physical pain because numerous previous studies had shown that there is greater empathic arousal in response to physical pain experienced by ingroup individuals compared to outgroup individuals<sup>32</sup>. However, Bruneau's study<sup>45</sup>, for instance, indicates larger effects of intergroup bias on emotional suffering relative to physical pain. Thus, subsequent studies should be expanded to include stimuli that present emotional suffering.

## Conclusion

The omnipresence of hate speech in human environments severely affects people's attitudes and behavior. Researchers of aggression and intergroup relations have often suggested that frequent exposure to hate speech might lead to prejudice, dehumanization, and lack of empathy toward outgroup members<sup>4,56</sup>. Yet the impairment of empathy among people exposed to hate speech has been proposed only theoretically<sup>4,56–58</sup>, and never directly empirically.

In the present experimental study, we aimed to directly tackle this question by analyzing the effects of hate speech exposure on empathy and/or mentalizing. The study suggests that immersion in a hateful environment leads to empathic numbing: people exposed to hate speech have limited ability to attribute the psychological perspective of others, regardless of their group membership.

This initial finding suggests that hate speech affects human functioning beyond intergroup relations. It is well established that after being exposed to hate speech against a specific target group (e.g., gay people), people start dehumanizing them<sup>56</sup> and their sensitivity to hate speech toward this group is reduced<sup>9</sup>. The present study shows that the effects of hate speech cannot be reduced to a single target group. Exposure to hate speech leads to a more general emotional numbing that affects perspective-taking not only toward certain target minority groups but also toward any other individual, including fellow ingroup members. This very basic psychological transformation is not only important from the perspective of human intergroup relations, but it also shows that hate speech poses a threat to any harmonious human interactions and everyday compassion.

## Data availability

Group-level whole-brain results maps are available to download from: <https://neurovault.org/collections/WGFVWQDF/>. The remaining set of data that support the findings of this study are available on request from the corresponding author.

Received: 22 August 2022; Accepted: 7 March 2023

Published online: 13 March 2023

## References

- Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M. & Räsänen, P. Exposure to online hate among young social media users. In *Soul of Society: A Focus on the Lives of Children and Youth* Vol. 18 (ed. Warehime, M. N.) 253–273 (Emerald Group Publishing Limited, 2014).
- Soral, W., Liu, J. & Bilewicz, M. Media of contempt: Social media consumption predicts normative acceptance of anti-Muslim hate speech and Islamoprejudice. *Int. J. Conflict Violence IJCV* <https://doi.org/10.4119/ijcv-3774> (2020).
- Brown, A. What is hate speech? Part 1: The myth of hate. *Law Philos.* **36**, 419–468 (2017).
- Bilewicz, M. & Soral, W. Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Polit. Psychol.* <https://doi.org/10.1111/pops.12670> (2020).
- Anderson, C. A. & Bushman, B. J. Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychol. Sci.* **12**, 353–359 (2001).
- Funk, J. B. Children's exposure to violent video games and desensitization to violence. *Child Adolesc. Psychiatr. Clin. N. Am.* **14**, 387–404 (2005).
- Krahé, B. & Möller, I. Longitudinal effects of media violence on aggression and empathy among German adolescents. *J. Appl. Dev. Psychol.* **31**, 401–409 (2010).
- Vossen, H. G. M., Piotrowski, J. T. & Valkenburg, P. M. The longitudinal relationship between media violence and empathy: Was it sympathy all along?. *Media Psychol.* **20**, 175–193 (2017).
- Soral, W., Bilewicz, M. & Winiewski, M. Exposure to hate speech increases prejudice through desensitization. *Aggress. Behav.* **44**, 136–146 (2018).
- Soral, W., Malinowska, K. & Bilewicz, M. The role of empathy in reducing hate speech proliferation. Two contact-based interventions in online and off-line settings. *Peace Conflict J. Peace Psychol.* <https://doi.org/10.1037/pac0000602> (2022).
- Decety, J. & Jackson, P. L. The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* **3**, 71–100 (2004).
- Lantos, D. & Molenberghs, P. The neuroscience of intergroup threat and violence. *Neurosci. Biobehav. Rev.* **131**, 77–87 (2021).
- Fourie, M. M., Stein, D. J., Solms, M., Gobodo-Madikizela, P. & Decety, J. Effects of early adversity and social discrimination on empathy for complex mental states: An fMRI investigation. *Sci. Rep.* **9**, 12959 (2019).

14. Fourie, M. M., Subramoney, S. & Madikizela, P. G. A less attractive feature of empathy: Intergroup empathy bias. In *Empathy—An Evidence-based Interdisciplinary Perspective* (ed. Kondo, M.) (InTech, 2017). <https://doi.org/10.5772/intechopen.69287>.
15. Bruneau, E. G., Cikara, M. & Saxe, R. Minding the gap: Narrative descriptions about mental states attenuate parochial empathy. *PLoS ONE* **10**, e0140838 (2015).
16. Bruneau, E. G., Cikara, M. & Saxe, R. Parochial empathy predicts reduced altruism and the endorsement of passive harm. *Soc. Psychol. Personal. Sci.* **8**, 934–942 (2017).
17. Zaki, J. & Cikara, M. Addressing empathic failures. *Curr. Dir. Psychol. Sci.* **24**, 471–476 (2015).
18. Kliemann, D. & Adolphs, R. The social neuroscience of mentalizing: Challenges and recommendations. *Curr. Opin. Psychol.* **24**, 1–6 (2018).
19. Harris, L. T. & Fiske, S. T. Dehumanized perception: A psychological means to facilitate atrocities, torture, and genocide?. *Z. Psychol.* **219**, 175–181 (2011).
20. Harris, L. T. & Fiske, S. T. Social neuroscience evidence for dehumanised perception. *Eur. Rev. Soc. Psychol.* **20**, 192–231 (2009).
21. Harris, L. T. & Fiske, S. T. Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychol. Sci.* **17**, 847–853 (2006).
22. Cameron, C. D. & Payne, B. K. Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *J. Pers. Soc. Psychol.* **100**, 1–15 (2011).
23. Cameron, C. D. *Compassion Collapse* Vol. 1 (Oxford University Press, 2017).
24. Figley, C. R. *Compassion Fatigue: Coping with Secondary Traumatic Stress Disorder in Those Who Treat the Traumatized* (Routledge, 2013). <https://doi.org/10.4324/9780203777381>.
25. Arioli, M., Cattaneo, Z., Ricciardi, E. & Canessa, N. Overlapping and specific neural correlates for empathizing, affective mentalizing, and cognitive mentalizing: A coordinate-based meta-analytic study. *Hum. Brain Mapp.* **42**, 4777–4804 (2021).
26. Dodel-Feder, D., Koster-Hale, J., Bedny, M. & Saxe, R. fMRI item analysis in a theory of mind task. *Neuroimage* **55**, 705–712 (2011).
27. Saxe, R. Uniquely human social cognition. *Curr. Opin. Neurobiol.* **16**, 235–239 (2006).
28. Kovács, A. M., Kühn, S., Gergely, G., Csibra, G. & Brass, M. Are all beliefs equal? Implicit belief attributions recruiting core brain regions of theory of mind. *PLoS ONE* **9**, e106558 (2014).
29. Bernhardt, B. C. & Singer, T. The neural basis of empathy. *Annu. Rev. Neurosci.* **35**, 1–23 (2012).
30. Fallon, N., Roberts, C. & Stancak, A. Shared and distinct functional networks for empathy and pain processing: A systematic review and meta-analysis of fMRI studies. *Soc. Cogn. Affect. Neurosci.* **15**, 709–723 (2020).
31. Shamay-Tsoory, S. The neuropsychology of empathy: evidence from lesion studies | Cairn.info. *Rev. Neuropsychol.* (2016).
32. Molenberghs, P. & Louis, W. R. Insights from fMRI studies into ingroup bias. *Front. Psychol.* **9**, 1868 (2018).
33. Xu, X., Zuo, X., Wang, X. & Han, S. Do you feel my pain? Racial group membership modulates empathic neural responses. *J. Neurosci.* **29**, 8525–8529 (2009).
34. Mathur, V. A., Harada, T., Lipke, T. & Chiao, J. Y. Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage* **51**, 1468–1475 (2010).
35. Winiewski, M., Hansen, K., Bilewicz, M., Soral, W., Świdorska, A. & Bulska, D. MOWA\_NIENAWISCI\_MOWA\_POGARDY\_INTERNET.pdf. *Fundacja Batorego/Centrum Badań nad Upżędzeniami UW, Warszawa* (2017).
36. Bilewicz, M., Soral, W., Marchlewska, M. & Winiewski, M. When authoritarians confront prejudice differential effects of SDO and RWA on support for hate-speech prohibition. *Polit. Psychol.* **38**, 87–99 (2017).
37. Pratto, F., Sidanius, J., Stallworth, L. M. & Malle, B. F. Social dominance orientation: A personality variable predicting social and political attitudes. *J. Pers. Soc. Psychol.* **67**, 741–763 (1994).
38. Bilewicz, M., Marchlewska, M., Soral, W. & Winiewski, M. (2014) Hate speech in Poland 2014. Summary of the national opinion poll. Stefan Batory Foundation.
39. Bruneau, E. G., Pluta, A. & Saxe, R. Distinct roles of the “shared pain” and “theory of mind” networks in processing others’ emotional suffering. *Neuropsychologia* **50**, 219–231 (2012).
40. Langner, O. et al. Presentation and validation of the Radboud faces database. *Cogn. Emot.* **24**, 1377–1388 (2010).
41. McFarquhar, M. et al. Multivariate and repeated measures (MRM): A new toolbox for dependent and multimodal group-level neuroimaging data. *Neuroimage* **132**, 373–389 (2016).
42. Xia, M., Wang, J. & He, Y. BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8**, e68910 (2013).
43. Wobbrock, J. O., Findlater, L., Gergely, D. & Higgins, J. J. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the 2011 annual conference on Human factors in computing systems—CHI ’11* 143 (ACM Press, 2011). <https://doi.org/10.1145/1978942.1978963>.
44. Timmers, I. et al. Is Empathy for pain unique in its neural correlates? A meta-analysis of neuroimaging studies of empathy. *Front. Behav. Neurosci.* **12**, 289 (2018).
45. Bruneau, E. G., Dufour, N. & Saxe, R. Social cognition in members of conflict groups: behavioural and neural responses in Arabs, Israelis and South Americans to each other’s misfortunes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 717–730 (2012).
46. Hudson, S. T. J., Cikara, M. & Sidanius, J. Preference for hierarchy is associated with reduced empathy and increased counter-empathy towards others, especially out-group targets. *J. Exp. Soc. Psychol.* **85**, 103871 (2019).
47. Koster-Hale, J. et al. Mentalizing regions represent distributed, continuous, and abstract dimensions of others’ beliefs. *Neuroimage* **161**, 9–18 (2017).
48. Saxe, R. & Powell, L. J. It’s the thought that counts: Specific brain regions for one component of theory of mind. *Psychol. Sci.* **17**, 692–699 (2006).
49. Young, L., Cushman, F., Hauser, M. & Saxe, R. The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 8235–8240 (2007).
50. Kim, J. J., Cunnington, R. & Kirby, J. N. The neurophysiological basis of compassion: An fMRI meta-analysis of compassion and its related neural processes. *Neurosci. Biobehav. Rev.* **108**, 112–123 (2020).
51. Anderson, C. A. & Dill, K. E. Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *J. Pers. Soc. Psychol.* **78**, 772–790 (2000).
52. Carnagey, N. L., Anderson, C. A. & Bushman, B. J. The effect of video game violence on physiological desensitization to real-life violence. *J. Exp. Soc. Psychol.* **43**, 489–496 (2007).
53. Olivo, D. et al. Neural responses of benefiting from the prosocial exchange: the effect of helping behavior. *Front. Psychol.* **12**, 606858 (2021).
54. Kühn, S. et al. Does playing violent video games cause aggression? A longitudinal intervention study. *Mol. Psychiatry* **24**, 1220–1234 (2019).
55. Szucs, D. & Ioannidis, J. P. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *Neuroimage* **221**, 117164 (2020).
56. Fasoli, F. et al. Not “just words”: Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *Eur. J. Soc. Psychol.* **46**, 237–248 (2016).
57. Murrow, G. B. & Murrow, R. A hypothetical neurological association between dehumanization and human rights abuses. *J. Law Biosci.* **2**, 336–364 (2015).

58. Roginsky, A. B. & Tsesis, A. Hate speech, volition, and neurology. *J. Law Biosci.* **3**, 174–177 (2016).

## Acknowledgements

This work has been funded by the Polish National Science Center Grant Sonata Bis awarded to Michał Bilewicz (Grant No. UMO-2017/26/E/HS6/00129). We thank Mikołaj Winiewski and Olga Kaminska for providing insightful discussion.

## Author contributions

A.P.: Conceptualization (lead), formal analysis (equal), investigation (supporting), methodology (equal), supervision (lead), writing—original draft preparation (lead); writing—review and editing (equal); J.M.: Formal Analysis (equal), investigation (lead), writing—original draft preparation (equal); J.W.: Data curation (lead), formal analysis (lead), methodology (equal), visualization (lead), writing—original draft preparation (supporting); writing—review and editing (equal); T.W.: Methodology (equal), project administration (lead), resources (lead); supervision (equal), writing—review and editing (supporting); W.S.: Formal analysis (equal), visualization (equal), writing—original draft preparation (equal); writing—review and editing (equal); M.B.: Funding acquisition (lead), supervision (equal), writing—original draft preparation (equal), writing—review and editing (equal).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-31146-1>.

**Correspondence** and requests for materials should be addressed to A.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023